A Generic Framework for Multiword Expressions Treatment: from Acquisition to Applications

Carlos Ramisch
Federal University of Rio Grande do Sul (Brazil)
GETALP — LIG, University of Grenoble (France)
ceramisch@inf.ufrgs.br

Abstract

This paper presents an open and flexible methodological framework for the automatic acquisition of multiword expressions (MWEs) from monolingual textual corpora. This research is motivated by the importance of MWEs for NLP applications. After briefly presenting the modules of the framework, the paper reports extrinsic evaluation results considering two applications: computer-aided lexicography and statistical machine translation. Both applications can benefit from automatic MWE acquisition and the expressions acquired automatically from corpora can both speed up and improve their quality. The promising results of previous and ongoing experiments encourage further investigation about the optimal way to integrate MWE treatment into these and many other applications.

SRC	*I* **paid** *my poor parents* **a visit**
MT	*J'ai* **payé** *mes pauvres parents* **une visite**
REF	*J'ai* **rendu visite** *à mes pauvres parents*
SRC	*Students pay* **an arm and a leg** *to park on campus*
MT	*Les étudiants paient* **un bras et une jambe** *pour se garer sur le campus*
REF	*Les étudiants paient* **les yeux de la tête** *pour se garer sur le campus*
SRC	*It shares the* **translation-invariance and homogeneity properties** *with the central moment*
MT	*Il partage la* **traduction-invariance et propriétés d'homogénéité** *avec le moment central*
REF	*Il partage les* **propriétés d'invariance par translation et d'homogénéité** *avec le moment central*

Table 1: Examples of SMT errors due to MWEs.

1 Introduction

Multiword expressions (MWEs) range over linguistic constructions such as idioms (*to pay an arm and a leg*), fixed phrases (*rock 'n' roll*) and noun compounds (*dry ice*). There is no unique and widely accepted definition for the term *multiword expression*. It can be an "arbitrary and recurrent word combination" (Smadja, 1993) or "a syntactic and semantic unit whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components" (Choueka, 1988) or simply an "idiosyncratic interpretation that crosses word boundaries (or spaces)" (Sag et al., 2002). MWEs lie in the fuzzy zone between lexicon and syntax, thus constituting a real challenge for NLP systems. In addition, they are very pervasive, occurring frequently in everyday language as well as in specialised communications. Some common properties of MWEs are:[1]

- **Arbitrariness**: sometimes valid constructions are not acceptable because people do not use them. Smadja (1993, p. 143–144) illustrates this by presenting 8 different ways of referring to the Dow Jones index, among which only 4 are used.
- **Institutionalisation**: MWEs are recurrent, as they correspond to conventional ways of saying things. Jackendoff (1997) estimates that they compose half of the entries of a speaker's lexicon, and Sag et al. (2002) point out that this may be an underestimate if we consider domain-specific MWEs.
- **Limited semantic variability**: MWEs do not undergo the same semantic compositionality rules as ordinary word combinations. This is expressed in terms of **(i) non-compositionality**, as the meaning of the whole expression often cannot be directly inferred from the meaning of the parts composing it, **(ii) non-substitutability**, as it is not possible to replace part of an MWE by a related (synonym/equivalent) word or construction, and **(iii) no word-for-word translation**.

[1] These are not binary yes/no flags, but values in a continuum going from flexible word combinations to prototypical fixed expressions.

Proceedings of the 2012 Student Research Workshop, pages 61–66,

Jeju, Republic of Korea, 8-14 July 2012. ©2012 Association for Computational Linguistics

- **Limited syntactic variability**: standard grammatical rules do not apply to MWEs. This can be expressed in terms of **(i) lexicalisation**, as one cannot list all MWEs in the lexicon (undergeneration) nor include them all in the grammar (overgeneration) and **(ii) extragrammaticality**, as MWEs are unpredictable and seem "weird" for a second language learner who only knows general rules.[2]
- **Heterogeneity**: MWEs are hard to define because they encompass a large amount of phenomena. Thus, NLP applications cannot use a unified approach and need to rely on some typology[3].

In this paper, I adopt the definition by Calzolari et al. (2002), who define MWEs as:

> different but related phenomena [which] can be described as a sequence[4] of words that acts as a single unit at some level of linguistic analysis.

This generic and intentionally vague definition can be narrowed down according to the application needs. For example, for the statistical machine translation (MT) system[5] used in the examples shown in Table 1, an MWE is any sequence of words which, when not translated as a unit, generates errors: ungrammatical or unnatural verbal constructions (sentence 1), awkward literal translations of idioms (sentence 2) and problems of lexical choice and word order in specialised texts (sentence 3). These examples illustrate the importance of correctly dealing with MWEs in MT applications and, more generally, MWEs can speed up and help remove ambiguities in many current NLP applications, for example:

- **Lexicography**: Church and Hanks (1990) used a lexicographic environment as their evaluation scenario, comparing manual and intuitive research with the automatic association ratio they proposed.
- **Word sense disambiguation**: MWEs tend to be less polysemous than simple words. Finlayson and Kulkarni (2011) exemplify that the word *world* has 9 senses in Wordnet 1.6, *record* has 14, but *world record* has only 1.
- **POS tagging and parsing**: recent work in parsing and POS tagging indicates that MWEs can help remove syntactic ambiguities (Seretan, 2008).
- **Information retrieval**: when MWEs like *pop star* are indexed as a unit, the accuracy of the system improves on multiword queries (Acosta et al., 2011).

2 Thesis contributions

Despite the importance of MWEs in several applications, they are often neglected in the design and construction of real-life systems. In 1993, Smadja pointed out that "…although disambiguation was originally considered as a performance task, the collocations retrieved have not been used for any specific computational task." Most of the recent and current research in the MWE community still focuses on MWE acquisition instead of integration of automatically acquired or manually compiled resources into applications. The main contribution of my thesis is that it represents a step toward the integration of automatically extracted MWEs into real-life applications. Concretely, my contributions can be classified in two categories: first, I propose a unified, open and flexible *methodological framework* (§ 3) for automatic MWE acquisition from corpora; and second, I am performing an intrinsic and extrinsic *evaluation of MWE acquisition* (§ 4), dissecting the influence of the different types of resources employed in the acquisition on the quality of the MWEs. The results of ongoing experiments are interesting but further work is needed to better understand the contributions of MWEs to the systems (§ 5).

Methodological Framework To date, there is no agreement on whether there is a single best method for MWE acquisition, or whether a different subset of methods works better for a given MWE type. Most of recent work on MWE treatment focuses on candidate extraction from preprocessed text (Seretan, 2008) and on the automatic filtering and ranking through association measures (Evert, 2004; Pecina, 2010), but few authors provide a whole picture of the MWE treatment pipeline. One of the advantages of the framework I propose is that it models the whole acquisition process with modular tasks that can be chained in several ways, each task having multiple available techniques. Therefore, it is highly customisable and allows for a large number of parameters to be tuned according to the target MWE types. Moreover, the techniques I have developed do not depend on a fixed length of candidate expression nor on adjacency assumptions, as the words in an expression might occur several words away. Thanks to this flexibility, this methodology can be easily applied to virtually any language, MWE type and domain, not strictly depending on a given formalism or tool[6]. Intuitively, for a given language, if some preprocessing tools like POS taggers and/or parsers are available, the results will be much better than running the methods on raw text. But since such tools are not available for all languages, the methodology was conceived to be applicable even in the absence of preprocessing.

[2] Examples of MWEs that breach standard grammatical rules include *kingdom come* and *by and large*.

[3] For example, Smadja (1993) classifies them according to syntactic function while Sag et al. (2002) classify them according to flexibility.

[4] Although they define MWEs as "sequences", assuming contiguity, we assume "sets" of words for greater generality.

[5] Automatic translations (MT) by Google (`http://translate.google.com/`) on 2012/02/18. Reference (REF) by native speaker.

[6] However, it is designed to deal with languages that use spaces to separate words. Thus, when working with Chinese, Japanese, or even with German compounds, some additional preprocessing is required.

Evaluation of MWE Acquisition Published results comparing MWE extraction techniques usually evaluate them on small controlled data sets using objective measures such as precision, recall and mean average precision (Schone and Jurafsky, 2001; Pearce, 2002; Evert and Krenn, 2005). On the one hand, the results of *intrinsic evaluation* are often vague or inconclusive: although they shed some light on the optimal parameters for the given scenario, they are hard to generalise and cannot be directly applied to other configurations. The quality of acquired MWEs as measured by objective criteria depends on the language, domain and type of the target construction, on corpus size and genre, on already available resources[7], on the applied filters, preprocessing steps, etc. On the other hand, *extrinsic evaluation* consists of inserting acquired MWEs into a real NLP application and evaluating the impact of this new data on the overall performance of the system. For instance, it may be easier to ask a human annotator to evaluate the output of an MT system than to ask whether a sequence of words constitutes an MWE. Thus, another original contribution of my thesis is application-oriented extrinsic evaluation of MWE acquisition on two study cases: computer-aided lexicography and statistical machine translation. My goal is to investigate (1) how much the MWEs impact on the application and (2) what is (are) the best way(s) of integrating them in the complex pipeline of the target application.

3 MWE Extraction

Among early work on developing methods for MWE identification, there is that of Smadja (1993). He proposed and developed a tool called Xtract, aimed at general-purpose collocation extraction from text using a combination of *n*-grams and a mutual information measure. On general-purpose texts, Xtract has a precision of around 80%. Since then, many advances have been made, either looking at MWEs in general (Dias, 2003), or focusing on specific MWE types, such as collocations, phrasal verbs and compound nouns. A popular type-independent approach to MWE identification is to use statistical association measures, which have been applied to the task with varying degrees of success (Evert and Krenn, 2005). One of the advantages of this approach is that it is language independent. This is particularly important since although work on MWEs in several languages has been reported, e.g. Dias (2003) for Portuguese and Evert and Krenn (2005) for German, work on English still seems to predominate.

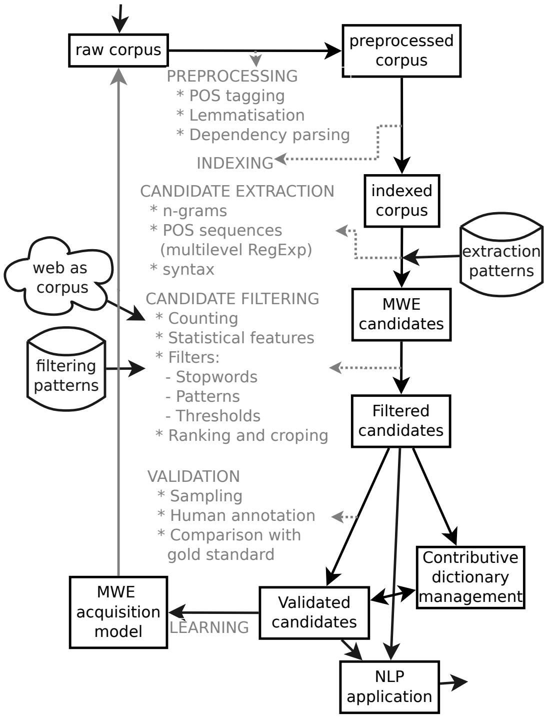

Figure 1: Framework for MWE acquisition from corpora

I propose a new framework called `mwetoolkit`, described in Figure 1, which integrates multiple techniques and covers the whole pipeline of MWE acquisition. One can preprocess a raw monolingual corpus, if tools are available for the target language, enriching it with POS tags, lemmas and dependency syntax. Then, based on expert linguistic knowledge, intuition, empiric observation and/or examples, one defines multilevel patterns in a formalism similar to regular expressions to describe the target MWEs. The application of these patterns on an indexed corpus generates a list of candidate MWEs. For filtering, a plethora of methods is available, ranging from simple frequency thresholds to stopword lists and sophisticated association measures. Finally, the resulting filtered candidates are either directly injected into an NLP application or further manually validated before application. An alternative use for the validated candidates is to train a machine learning model which can be applied on new corpora in order to automatically identify and extract MWEs based on the characteristics of the previously acquired ones. For further details, please refer to the website of the framework[8] and to previous publications (Ramisch et al., 2010a; Ramisch et al., 2010b).

4 Application-oriented evaluation

In this section, I present summarised results of extrinsic quantitative and qualitative evaluation of the framework for MWE acquisition propose in § 3. The target applications are computer-aided lexicography (§ 4.1) and statistical machine translation (§ 4.2).

[7] It is useless to acquire MWEs already present in the dictionary.

[8] `http://mwetoolkit.sf.net`

Language	Type	Corpus (words)	Candidates	Final MWEs	Publication
English	PV	Europarl (13M)	5.3K	875	(Ramisch et al., 2012)
French	NC	Europarl (14.5M)	104K	3,746	(Ramisch et al., 2012)
Greek	NC	Europarl (26M)	25K	815	(Linardaki et al., 2010)
Portuguese	CP	PLN-BR-FULL (29M)	407K	773	(Duran et al., 2011)

Table 2: MWE acquisition applied to lexicography

4.1 Computer-aided Lexicography

In this evaluation, I collaborated with colleagues who are experienced linguists and lexicographers, in order to create new lexical resources containing MWEs. The languages of the resources are English, French, Greek and Portuguese. Table 2 summarises the outcomes of each evaluation. The created data sets are freely available.[9, 10]

We extracted English phrasal verbs (PVs) from the English portion of the Europarl corpus[11]. We considered a PV as being formed by a verb (except *to be* and *to have*) followed by a prepositional particle[12] not further than 5 words after it[13] This resulted in 5,302 phrasal verb candidates occurring more than once in the corpus, from which 875 were automatically identified as true PVs and the others are currently under manual validation. Analogously, the French noun compounds (NCs) were extracted from Europarl using the following pattern: a noun followed by either an adjective or a prepositional complement[14]. After filtering out candidates that occur once in the corpus, we obtained 3,746 MWE candidates and part of the remaining candidates will be manually analysed in the future.

For Greek, in particular, considerable work has been done to study the linguistic properties of MWEs, but computational approaches are still limited (Fotopoulou et al., 2008). In our experiments, we extracted from the POS-tagged Greek part of the Europarl corpus words matching the following patterns: adjective-noun, noun-noun, noun-determiner-noun, noun-preposition-noun, preposition-noun-noun, noun-adjective-noun and noun-conjunction-noun. The candidates were counted in two corpora and annotated with four association measures, and the top 150 according to each measure where annotated by three native speakers, that is, each annotator judged around 1,200 candidates and in the end the annotations were joined, creating a lexicon with 815 Greek nominal MWEs.

Finally, the goal of the work with Portuguese complex predicates (CPs) was to perform a qualitative analysis of these constructions. Therefore, we POS-tagged the PLN-BR-FULL corpus[15] and extracted sequences of words matching the patterns: verb-[determiner]-noun-preposition, verb-preposition-noun, verb-[preposition/determiner]-adverb and verb-adjective. The extraction process resulted in a list of 407,014 candidates which were further filtered using statistical association measures. Thus, an expert human annotator manually validated 12,545 candidates from which 699 were annotated as compositional verbal expressions and 74 as idiomatic verbal expressions. Afterwards, a fine-grained analysis of each extraction pattern was conducted with the goal of finding correlations between syntactic flexibility and semantic properties such as compositionality.

4.2 Statistical Machine Translation (SMT)

Incorporating even simple treatments for MWEs in SMT systems can improve translation quality. For instance, Carpuat and Diab (2010) adopt two complementary strategies for integrating MWEs: a static strategy of single-tokenisation that treats MWEs as word-with-spaces and a dynamic strategy that adds a count for the number of MWEs in the source phrase. They found that both strategies result in improvement of translation quality, which suggests that SMT phrases alone do not model all MWE information. Morin and Daille (2010) obtained an improvement of 33% in the French–Japanese translation of MWEs with a morphologically-based compositional method for backing-off when there is not enough data in a dictionary to translate a MWE (e.g. *chronic fatigue syndrome* decomposed as *[chronic fatigue] [syndrome]*, *[chronic] [fatigue syndrome]* or *[chronic] [fatigue] [syndrome]*). For translating from and to morphologically rich languages like German, where a compound is in fact a single token formed through concatenation, Stymne (2011) splits the compound into its single word components prior to translation and then applies some post-processing, like the reordering or merging of the components, after translation. She obtains improvements in BLEU from 21.63 to 22.12 in English–Swedish and from 19.31 to 19.73 in English–German.

[9] `http://multiword.sourceforge.net/PHITE.php?sitesig=FILES&page=FILES_20_Data_Sets`

[10] `http://www.inf.ufrgs.br/~ceramisch/?page=downloads/mwecompare`

[11] `http://statmt.org/europarl`

[12] *up, off, down, back, away, in, on.*

[13] Even though the particle might occur further than 5 positions away, such cases are sufficiently rare to be ignored in this experiment.

[14] Prepositions *de*, *à* and *en* followed by optionally determined noun.

[15] `www.nilc.icmc.usp.br/plnbr`

	% Good	% Acceptable	% Incorrect
Baseline	0.53	0.36	0.11
TOK	0.55	0.29	0.16
PV?	0.50	0.39	0.11
PART	0.53	0.36	0.11
VERB	0.53	0.36	0.11
BILEX	0.50	0.29	0.20

Table 3: Evaluation of translation of phrasal verbs in test set.

In the current experiments, a standard non factored phrase-based SMT system was built using the open-source Moses toolkit with parameters similar to those of the baseline system for the 2011 WMT campaign. [16]. For training, we used the English–Portuguese Europarl v6 (EP) corpus, with 1.7M sentences and around 50M words. The training data contains the first 200K sentences tokenized and lowercased, resulting in 152,235 parallel sentences and around 3.1M words. The whole Portuguese corpus was used as training data for 5-gram language model built with SRILM. Phrasal verbs were automatically identified using the jMWE tool and a dictionary of PVs. We compared the following five strategies for the integration of automatically identified phrasal verbs in the system:

- TOK: before translation, rearrange the verb and the particle in a joint configuration and transform them into a single token with underscore (e.g. *call him up* into *call_up him*).
- PV?: add a binary feature to each bi-phrase indicating whether a source phrasal verb has been detected in it or not.
- PART: replace the particle by the one most frequently used with the target verb, using a web-based language model with a symmetric windows of 1 to 5 words around the particle.
- VERB: modify the form of the Portuguese verb (gerund or infinitive), according to the form detected on the English side.
- BILEX (or *bilingual lexicon*): augment the phrase table of the baseline system with 179,133 new bilingual phrases from an English–Portuguese phrasal verb lexicon.

Table 3 shows the preliminary results of a human evaluation performed on a test set of 100 sentences. The sentences were inspected and we verified that, while some translations improve with the integration strategies, others are degraded. No absolute improvement was observed, but we believe that this is due to the fact that our evaluation needs to consider more fine-grained classes of phrasal verbs instead of mixing them all in the same test set. Additionally, we would need to annotate more data in order to obtain more representative results. These hypotheses motivate us to continue our investigation in order to obtain a deeper understanding the impact of each integration strategy on each step of the SMT system.

5 Future Experiments and Perspectives

In this paper, I described an open framework for the automatic acquisition of MWEs from corpora. What distinguishes it from related work is that it provides an integrated environment covering the whole acquisition pipeline. For each module, there are multiple available techniques which are flexible, portable and can be combined in several ways. The usefulness of the framework is then presented in terms of extrinsic application-based evaluation. I presented summarised results of ongoing experiments in computer-aided lexicography and in SMT.

Although our results are promising, the experiments on SMT need further investigation. I am currently applying syntax-based identification and analysing word alignment and translation table entries for a set of prototypical MWEs, in order to obtain a better understanding of the impact of each integration strategy on the system. Moreover, I would like to pursue previous experiments on bilingual MWE acquisition from parallel and comparable resources. Finally, I would like to experiment on MWE simplification (e.g. replacing a multiword verb like *go back* by its simplex form *regress*) as preprocessing for SMT, in order to improve translation quality by making the source language look more like the target language.As these improvements depend in the MT paradigm, I would also like to evaluate strategies for the integration of verbal MWEs in expert MT systems.

In spite of a large amount of work in the area, the treatment of MWEs in NLP applications is still an open and challenging problem. This is not surprising, given their complex and heterogeneous behaviour (Sag et al., 2002). At the beginning of the 2000's, Schone and Jurafsky (2001) asked whether the identification of MWEs was a solved problem, and the answer that paper gave was 'no, it is not'. The MWE workshop series have shown that this is still the case, listing several challenges in MWE treatment like lexical representation and application-oriented evaluation. Therefore, I believe that my thesis will be a significant step toward the full integration of MWE treatment in NLP applications, but there is still a long road to go.

Acknowledgements

This work was partly funded by the CAMELEON project (CAPES–COFECUB 707-11) and by a Ph.D. grant from the French Ministry for Higher Education and Research. I would

[16] www.statmt.org/wmt11/baseline.html

like to thank my supervisors Aline Villavicencio and Christian Boitet, as well as the colleagues who contributed to this work: Evita Linardaki, Valia Kordoni, Magali Sanchez Duran and Vitor De Araujo.

References

Otavio Acosta, Aline Villavicencio, and Viviane Moreira. 2011. Identification and treatment of multiword expressions applied to information retrieval. In Valia Kordoni, Carlos Ramisch, and Aline Villavicencio, editors, *Proc.of the ALC Workshop on MWEs: from Parsing and Generation to the Real World (MWE 2011)*, pages 101–109, Portland, OR, USA, Jun. ACL.

Nicoleta Calzolari, Charles Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine Macleod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proc. of the Third LREC (LREC 2002)*, pages 1934–1940, Las Palmas, Canary Islands, Spain, May. ELRA.

Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Proc. of HLT: The 2010 Annual Conf. of the NAACL (NAACL 2003)*, pages 242–245, Los Angeles, California, Jun. ACL.

Yaacov Choueka. 1988. Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *RIAO'88*, pages 609–624.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms mutual information, and lexicography. *Comp. Ling.*, 16(1):22–29.

Gaël Dias. 2003. Multiword unit hybrid extraction. In Francis Bond, Anna Korhonen, Diana McCarthy, and Aline Villavicencio, editors, *Proc. of the ACL Workshop on MWEs: Analysis, Acquisition and Treatment (MWE 2003)*, pages 41–48, Sapporo, Japan, Jul. ACL.

Magali Sanches Duran, Carlos Ramisch, Sandra Maria Aluísio, and Aline Villavicencio. 2011. Identifying and analyzing brazilian portuguese complex predicates. In Valia Kordoni, Carlos Ramisch, and Aline Villavicencio, editors, *Proc.of the ALC Workshop on MWEs: from Parsing and Generation to the Real World (MWE 2011)*, pages 74–82, Portland, OR, USA, Jun. ACL.

Stefan Evert and Brigitte Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. *Comp. Speech & Lang. Special issue on MWEs*, 19(4):450–466.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, Stuttgart, Germany.

Mark Finlayson and Nidhi Kulkarni. 2011. Detecting multiword expressions improves word sense disambiguation. In Valia Kordoni, Carlos Ramisch, and Aline Villavicencio, editors, *Proc.of the ALC Workshop on MWEs: from Parsing and Generation to the Real World (MWE 2011)*, pages 20–24, Portland, OR, USA, Jun. ACL.

Aggeliki Fotopoulou, Giorgos Giannopoulos, Maria Zourari, and Marianna Mini. 2008. Automatic recognition and extraction of multiword nominal expressions from corpora (in greek). In *Proceedings of the 29th Annual Meeting, Department of Linguistics*, Aristotle University of Thessaloniki, Greece.

Ray Jackendoff. 1997. Twistin' the night away. *Language*, 73:534–559.

Evita Linardaki, Carlos Ramisch, Aline Villavicencio, and Aggeliki Fotopoulou. 2010. Towards the construction of language resources for greek multiword expressions: Extraction and evaluation. In Stelios Piperidis, Milena Slavcheva, and Cristina Vertan, editors, *Proc. of the LREC Workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages*, pages 31–40, Valetta, Malta. May.

Emmanuel Morin and Béatrice Daille. 2010. Compositionality and lexical alignment of multi-word terms. *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing*, 44(1-2):79–95, Apr.

Darren Pearce. 2002. A comparative evaluation of collocation extraction techniques. In *Proc. of the Third LREC (LREC 2002)*, Las Palmas, Canary Islands, Spain, May. ELRA.

Pavel Pecina. 2010. Lexical association measures and collocation extraction. *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing*, 44(1-2):137–158, Apr.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010a. Multiword expressions in the wild? the mwetoolkit comes in handy. In Yang Liu and Ting Liu, editors, *Proc. of the 23rd COLING (COLING 2010) — Demonstrations*, pages 57–60, Beijing, China, Aug. The Coling 2010 Organizing Committee.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010b. mwetoolkit: a framework for multiword expression identification. In *Proc. of the Seventh LREC (LREC 2010)*, Malta, May. ELRA.

Carlos Ramisch, Vitor De Araujo, and Aline Villavicencio. 2012. A broad evaluation of techniques for automatic acquisition of multiword expressions. In *Proc. of the ACL 2012 SRW*, Jeju, Republic of Korea, Jul. ACL.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proc. of the 3rd CICLing (CICLing-2002)*, volume 2276/2010 of *LNCS*, pages 1–15, Mexico City, Mexico, Feb. Springer.

Patrick Schone and Daniel Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In Lillian Lee and Donna Harman, editors, *Proc. of the 2001 EMNLP (EMNLP 2001)*, pages 100–108, Pittsburgh, PA USA, Jun. ACL.

Violeta Seretan. 2008. *Collocation extraction based on syntactic parsing*. Ph.D. thesis, University of Geneva, Geneva, Switzerland.

Frank A. Smadja. 1993. Retrieving collocations from text: Xtract. *Comp. Ling.*, 19(1):143–177.

Sara Stymne. 2011. Pre- and postprocessing for statistical machine translation into germanic languages. In *Proc. of the ACL 2011 SRW*, pages 12–17, Portland, OR, USA, Jun. ACL.